DOCUMENT RESUME

ED 441 415 IR 057 686

AUTHOR	Byrum, John D.
TITLE	ISO 639-1 and ISO 639-2: International Standards for Language Codes. ISO 15924: International Standard for Names of Scripts.
PUB DATE	1999-08-00
NOTE	6p.; In: IFLA Council and General Conference. Conference Programme and Proceedings (65th, Bangkok, Thailand, August 20-28, 1999); see IR 057 674.
AVAILABLE FROM	For full text: http://www.ifla.org/IV/ifla65/papers/099-155e.htm.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Alphabets; Cataloging; *Coding; *Languages; *Standards
IDENTIFIERS	International Organization for Standardization; *International Standards

ABSTRACT

This paper describes two international standards for the representation of the names of languages. The first (ISO 639-1), published in 1988, provides two-letter codes for 136 languages and was produced primarily to meet terminological needs. The second (ISO 639-2) appeared in late 1998 and includes three-letter codes for 460 languages. This list addresses terminological needs, as well as bibliographic applications. For this reason, ISO 639-2 is covered in detail. Its features are explained, and principles and policies used for development of this code list are presented. Additionally, the governance mechanism established to maintain ISO 639-1 and ISO 639-2 are described. Also presented is a brief summary regarding a project in progress to provide codes for names of scripts that will result in publication of ISO 15924. The paper concludes that the emergence of an international standard for language codes and of the developing international standard for script codes is a major contribution to Universal Bibliographic Control, as these code lists enable important information regarding the nature of publications represented by records to be communicated and shared unambiguously, efficiently, and internationally. (Author/MES)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 415

IFLANET
International Federation of Library Associations and Institutions
Search Contacts
Annual Conference

65th IFLA Council and General Conference

Bangkok, Thailand,
August 20 - August 28, 1999

Conference Proceedings

Code Number: 099-155(WS)-E
Division Number: IV
Professional Group: Cataloguing: Workshop
Joint Meeting with: -
Meeting Number: 155
Simultaneous Interpretation: *No*

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
A.L. Van Wesermael
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ISO 639-1 and ISO 639-2: International Standards for Language Codes. ISO 15924: International Standard for names of scripts

John D. Byrum
Library of Congress
Washington DC, USA

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
- This document has been reproduced as received from the person or organization originating it
- Minor changes have been made to improve reproduction quality
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

Abstract

The author describes two international standards for the representation of the names of languages. The first (ISO-639-[1]) published in 1988 provides two-letter codes for 136 languages and was produced primarily to meet the terminological needs. The second (ISO 639-2) appeared in late 1998 and includes three-letter codes for 460 languages. This list addresses terminological needs but also for bibliographic applications. For this reason, 639-2 is covered in detail. Its features are explained, and principles and policies used for development of this code list are presented. Additionally, the author describes the governance mechanism established to maintain ISO 639-[1] and ISO 638-2. Also presented is a brief summary regarding a project in progress to provide codes for names of scripts and when completed to result in publication of ISO 15924. The paper concludes that " the emergence of an international standard for language codes and of the developing international standard for script codes is a major contribution to Universal Bibliographic Control as these code lists enable of important information regarding the nature of publications represented by records to be communicated and shared unambiguously, efficiently, and internationally."

Paper

The International Organization for Standardization (ISO) has long been interested in codes for the representation of names of languages. Work on a project to prepare a standard for two-letter codes (hereafter sometimes referred to as "alpha-2" codes) began several decades

IR057686
BEST COPY AVAILABLE

ago, although the publication of ISO 639 (hereafter referred to as ISO 639-1) did not occur until 1988[1]. In that year, work began on the production of a standard for three-letter codes (hereafter sometimes referred to as "alpha-3" codes), but it required another decade of work for the publication of ISO 639-2: *Codes for the Representation of Names of Languages: Alpha-3 Codes*[2] . Meanwhile, an effort was initiated in the mid-1990's on the revision of ISO 639-1; that project has yet to produce a Draft International Standard (DIS).

ISO's Technical Committee 37 (Terminology)/ Subcommittee 3 (Layout of Vocabularies) was responsible for ISO 639-1. As a result, this code list was devised primarily for use in terminology, lexicography and linguistics. ISO 639-1 lists 136 codes for as many language names. The alpha-2 code set was devised for practical use for most of the major languages of the world that are most frequently represented in the total body of the world's literature. Additional language codes are created when it becomes apparent that a significant body of literature in a particular language exists. The individual codes are based on the original name of the language if in Latin spelling or converted to the Latin script, except in a few cases where the appropriate national standardization organizations requested codes based on the English form of the language names. For example, the code for Japanese in ISO 639-1 is "ja".

According to the introduction for the alpha-2 language code standard, the terminological and linguistic usages most likely for such codes are: (1) to indicate the language used, for example, at heads of documents or in bibliographies and (2) to indicate the language to which a term belongs, for example, in documents on terminology, vocabularies, dictionaries, or in multilingual alphabetically ordered list of words. This standard does not directly mention bibliographic applications for language codes.

ISO 639-2 was the result of a joint undertaking in which TC37/SC 2 invited participation of representatives from ISO's Technical Committee 46 (Information and Documentation) Subcommittee 4 (Computer Applications in Information and Documentation). Consequently, this standard not only intends to provide for the linguistic applications mentioned above, but in the context of a much larger pool of language names; but ISO 639-2 also recognizes the usages for language codes for use by libraries, information services, and publishers to indicate language in the exchange of information, especially in computerized systems.

In the Introduction to ISO 639-2, the particular usages have been substantially broadened. In addition to recognizing the usages for terminological and linguistic purposes cited in ISO 639-1, the alpha-3 list notes that language codes provided for communication of bibliographic information. Such usages include indication of the languages in which documents are or have been written or recorded -- for example in UNIMARC Format, Field 101 to designate the Language of the item -- and indication of the languages in which document-handling records (order records, bibliographic records, and the like) have been created -- for example in the UNIMARC Format, Field 100 Language of cataloguing; positions 22-24.

From the viewpoint of the topic of this Workshop -- Universal Bibliographic Control in the Multilingual Environment -- ISO 639-2 with its bibliographic emphasis and greater pool of language codes is much more likely to meet the needs of those who create and consume descriptions of documents of all types than is ISO 639-1. For that reason the remainder of this presentation will focus on the alpha-3 code list.

ISO 639-2 represents all languages contained in ISO 639-1 and in addition many other languages as well as several language groups and some codes for special purposes. Currently, the languages listed in ISO 639-1 are a subset of the languages listed in ISO 639-2; every language code in the two-letter code set has a corresponding language code in the alpha-3 list, but of course vice versa. There are more than 460 codes contained in ISO 639-2. (Languages designed exclusively for machine use, such as computer programming languages, are not included in either code list.)

The Joint Working Group (JWG) which created ISO 639-2 decided early in the project to

make the codes in ISO 639-2 consistent with those in ISO 639-1 to the extent that it was practical to do so. However, in the development of the standard there was considerable difficulty over the choice of codes, since the bibliographic community had a well established list (based on the MARC 21[3] language code list) that was not always compatible with ISO 639-1. As a necessary compromise between the terminology community and the bibliographic community (which has used its codes for many years in hundreds of millions of bibliographic records), the JWG agreed to standardize two sets of codes, one for bibliographic applications (ISO 639-2/B) and one for terminology applications (ISO 639-2/T). The two sets are different in 23 language codes.

Code set B provides for bibliographic applications which largely require unique recognition of individual languages and language groups and do not depend necessarily on language names, as they are not necessarily intended to be an abbreviation for the language. Given the extensive use of the existing MARC 21 language codes in bibliographic records, the approach represented by that MARC list was largely adopted for Code set B. Thus, for the bibliographic list, the JWG established the following criteria for selecting the form of a language code - usually (but not invariably) in this order:

- preference of the countries using the language
- established usage of codes in national and international bibliographic databases, and
- the vernacular or English form of the language.

Since ISO 639-2 intends to provide for terminological needs as well, code set T was based on:

- the vernacular form of the language, or
- preference of the countries using the language.

Despite differences in criteria for Code set B and Code set T, to reiterate, there are only 23 language codes among the more than 460 included which are not identical in the two sets. In addition, the JWG agreed that future development of language codes should be based whenever possible on the vernacular form of the language, unless another language code is requested by the country or countries using the language. Narrowing the differences between Code sets B and T consumed a great part of the 10 years that went into production of ISO 639-2 and at many points it proved necessary to remind the JWG that the purpose of standardized lists of language codes is not to standardize the name of the language represented by the codes but to standardize the symbols. It is important to recognize that the representatives of the bibliographic community on the JWG made many concessions on behalf of compromise to bring the project to a successful conclusion. As a result, some 25 MARC 21 language codes will need to be changed, 33 new language codes will be added, and one will be made obsolete. The impact of so many changes on large bibliographic databases is a cause for concern, even though the codes to be changed represent languages which might be considered relatively minor depending on the contents of one's library collection. However, for the future, it is expected that the MARC list and ISO 639-2 will remain compatible.

Given the co-existence of the two alpha-3 code sets, whichever one is selected must be used in its entirety, and the choice of the set used must be made clear by exchanging partners prior to information interchange. There is no option available to use one or a few codes from one of the sets while using codes from the other sets. The JWG also agreed to include a policy statement to the effect that codes may only be changed for compelling reasons and after a change is made, any previous codes can not be reused for at least five years. Another principle which applies specifically to the Bibliographic set is that codes in ISO 639-2/B will not be changed in order to reduce database maintenance work should the name of a language change, as happened, for example, when Gallegan changed to Galician and Langue d'oc changed to Occitan.

There are some special features incorporated within ISO 639-2 which are not present in the alpha-2 code list. One of these is provision of "collective language codes" which are used in

cases where the body of literature is relatively sparse. To be considered for its own language code, a body of literature equaling at least 50 unique titles must be held by a single institution or 50 held by five agencies among them; the total may include titles in any format, however, not just those which have been published as printed works. The particular languages included in each collective code are not specified in ISO 639-2 as they are in the MARC 21 list, as a result of a decision by the Joint Working Group. Another special feature of the alpha-3 list is inclusion of a code (mul) to be used in records for works which include parts that are in multiple languages and a code (und) to be used in cases where it is necessary to provide a language code but the name of the language is not known to the person creating the record.

A single language code is normally provided for a language even though the language is written in more than one script - for example, in the case of Sindhi, which is written in Arabic, Gurmukhi, and Devangari scripts, or Somali, written in Arabic and Roman scripts. As the single exception, separate codes are provided for Croatian and Serbian, although most experts feel that this is a situation where the same language is employed but by some users in the Roman alphabet and by others in Cyrillic.

Usually, all dialects of a language are represented by the code for the language but in a few cases, mostly as a result of historical circumstances rather than principle, codes are present for dialects - for example, in the case of Awadhi which is a dialect of Hindi. ISO 639-2 recognizes that in some cases agencies will want to provide codes for dialects which are not present in the standard; thus, the codes qaa through qtx have been reserved for local use. Another case where local codes might be wanted would be to provide codes for those ancient languages which in ISO 639-2 do not have their own codes. But the standard warns that records with codes assigned from those reserved for local use should not be exchanged internationally, as the local codes will differ from institution to institution.

Maintaining an international standard, of course, is an essential activity to enable confidence that it addresses changing circumstances and requirements. In the case of ISO 639, two Registration Authorities have been appointed. The Registration Authority for the alpha-2 list is Infoterm, which is located in Vienna, Austria, while that for the alpha-3 list is the Library of Congress in Washington, D.C., USA. Both Registration Authorities are responsible for receiving and evaluating proposals for new or changed language codes. As already mentioned, requests for additional codes not now represented among the individual languages included in ISO 639-2 must be supported by evidence of 50 titles. When a request for a new code has been rejected, the code requested may be reserved for use of the applicant and other possible users. In every case, the Registration Authorities recommend actions to a Joint Advisory Committee (JAC) which oversees the standard as a whole. The JAC has equal representation from TC 37 and 46 constituents, and the chair rotates every two years between Infoterm and the Library of Congress. In making decisions, the JAC must be unanimous in passing proposals on the first vote; if that is not possible then a second vote is required and at least five positive votes are needed for a proposal to pass. The equality of representation and the stringent requirement for consensus will help to ensure that the future development of ISO 639-1 is well considered.

The Joint Advisory Committee is expected to meet in October 1999 to discuss technical issues which arose during the comment periods when 639-2 advanced to Draft International Standard status. A more general issue which needs to be discussed is the future relationship between the alpha-3 language codes and those included in the alpha-2 list comprising ISO 639-1 which is now undergoing independent revision.

Another ISO project of relevance to the theme of this workshop is the effort currently in progress by TC46/SC2 (Conversion of written languages) to produce a code list for the representation of names of scripts.[4] These codes like the language code described above are intended for use in terminology, lexicography, and linguistics as well as for any application requiring the expression of scripts in coded form, including of course machine manipulation for bibliographic purposes. This proposed standard offers not one but three codes for each script name included: (1) a two-letter code and (2) a three-letter code usually created from the

original script name in the language commonly used for it, transliterated or transcribed into Latin letters, as well as (3) a numeric version designed "to provide some measure of mnemonicity to the codes used." For the numeric codes ranges of numbers have been designated to cover the nature of the script: for example 000-099 are for hieroglyphic and cuneiform scripts, 100-199 for right-to-left scripts, 200-299 for left-to right scripts and so on. The numbers 700-899 are unassigned while 900-999 are reserved "for private use, aliases for multiple scripts and special codes." The alphabetic script codes are described from ISO 639-1 and ISO 639-2, with no particular preference given to either Terminological or Bibliographic alternatives present in the latter standard. Included in the current draft are codes for about 95-100 scripts and aliases. Once adopted as an international standard, ISO 15924 will be maintained by a Registration Authority yet to be designated.

In conclusion, the emergence of an international standard for language codes and of the developing international standard for script codes is a major contribution to Universal Bibliographic Control as these code lists enable important information regarding the nature of publications represented by records to be communicated and shared unambiguously, efficiently, and internationally.

Endnotes:

1. ISO 639: Code for the Representation of Names of Languages. 1st Edition. Geneva: International Standardization Organization, 1988. 17 p.

2. ISO 639-2: Codes for the Representation of Names of Languages: Alpha-3. 1st Edition. Geneva: International Standardization Organization, 1998. 66 p.

3. MARC 21 is the name of the recently harmonized USMARC and CAN/MARC formats, published in 1999.

4. CD for ISO 15:924: Code for the Representation of Names of Scripts. Committee draft, dated July 9, 1998. 18 p.

Latest Revision: *June 22, 1999*

Copyright © 1995-1999
International Federation of Library Associations and Institutions
www.ifla.org